Rutgers University, 16:954:597 Data Wrangling and Husbandry, Spring 2022

Project Proposal

Rajesh Bhat-rmb317   Rutu Desai-rmd228   Shubham Kokane-ssk203

Professor - Tirthankar Dasgupta

March 14, 2022

---

1. Project Goal

- The main aim of the project is to *analyse the sentiments of the comments from different YouTube Comments* and to *find the topic* based on the classification of the comments using *Latent Dirichlet Allocation (LDA)* method and try to present the sentiment score for each video by establishing a colour grading for each sentiment on public opinions.

2. Motivation

- We know that data is generated tremendously every second of the day and YouTube being one of the most favourite video streaming platform which has a lot of data generated on each video which can have different types of context (semantically). To find if the comments are acceptable or not it's very difficult nowadays so to find the Topic from comments and finding the comments which are good and bad so that bad comments can be filtered is really necessary.

3. Dataset

- We will scrape the data from different genre YouTube comment sections and append it to a .csv file for further analysis.

4. Procedure

1. We will start by scraping data from YouTube comments of different genres.
2. Use different methodologies to clean the data, for instance using regex to find specific patterns, removing stop words.
3. Use Data Visualizations to show different graphs and charts.
4. Will then compare both methods of LDA (VEM and Gibbs) and try to compare each model with different 'k' values.
5. Try to find/guess topics from the model created above.
6. Give out the sentiment score of each video.
7. Colour grading will be established on public sentiments.